A quick review:

After reading the paper by Keiser et al. (2007) [please give the full citation in Bibliography at the end of the summary.] describing the Similarity ensemble approach (SEA) it appeared that in the process of creating the statistical model, the reasoning behind some of the steps was lacking or not provided. Thus, we decided to derive a statistically robust model.

Methods:

All ligands and drugs fingerprints [please define ligand, drug fingerprint and the difference between them. Maybe give an example.], target (protein) ligand [what is “target ligand”?] groups were provided by Keiser from a paper published in 2012 testing SEA predictions in-vitro.

Sampling process:

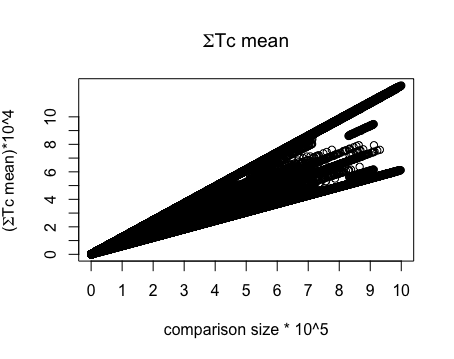
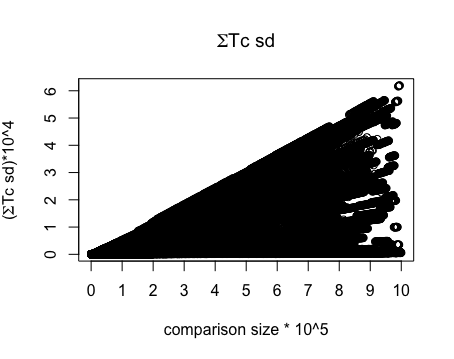
1. Ligand sets of sizes ranging from 1 to 1,000 were randomly selected from a pool of 163,547 ligands.
2. For all pairs of ligands, where one is from set A and the other from set B, the Tanimoto coefficient score (Tc; a measure of the chemical similarity between A and B. [Please give the range. What are the values with maximal and minimal similarity.]) was calculated. The sum of Tc was saved, so a tuple was created : (, , ).

The sampling process was repeated 10 times.

Analysis done:

In SEA (Keiser et al. 2007) sampling was done using “across logarithmic set size intervals in the range of 10 to 1,000 molecules”. In that paper there were no sets of a single ligand in the sampling process. However, the same statistical model was also used in the next publication (Keiser et al. 2012) [add to bibliography] to predict the statistical significance when a single ligand is checked against another set of ligands which sometimes results in a comparison size less than 100 [please define ‘comparison size’.]. Thus, we examined if the same results would be achieved when using set sizes ranging from 1 to 1,000 (using set size interval of 1).

[Please use active voice when describing your work. “We calculated”, etc. It is much more simple to write and read in active voice.] Using the data sampled, a mapping between and was calculated in order to replicated the sampling data produced by Keiser et al. for creating the SEA statistical model.

The results are shown in the following two figures [Please number the figures and refer to them in the text using the numbers. This would be Figure 1.]: 

To the left is a plot showing the mean of for each comparison size (which equals in previous notation). When fitting a linear model to mean the model has an R-squared value of 0.8421.

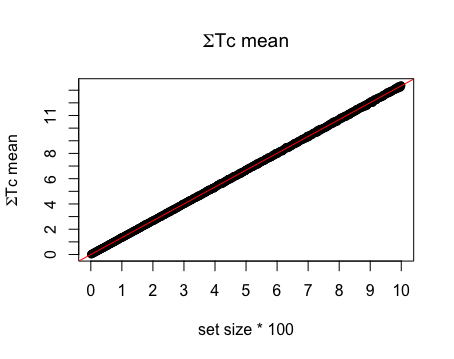
To the right is a plot showing the standard deviation (SD) of for each comparison size.

When fitting a linear model to SD the model has an R-squared value of 0.1325. It is easy to see that trying to fit any model to that data would not produce usable results. (In the paper describing SEA [Keiser 2007?] a nonlinear fit was computed). [The plot shows multiple linear lines. Could you suggest why? If not, please report this observation and leave it as open question.]

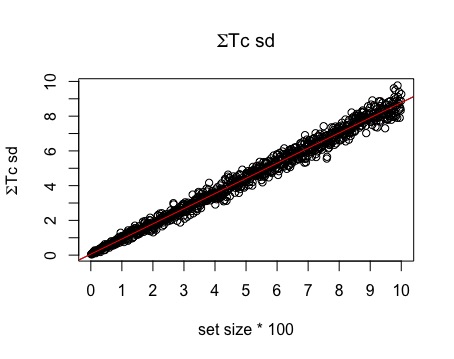
We examine the possibility that the data above corresponds to a mixture. Specifically, that the results of comparisons of the same size may differ based on the sizes of the sets used. For example, a comparison of size 400 can be obtained by comparing one ligand to a set of 400 ligands, as well as by comparing two sets of 20 ligands in each. We examine if the results of the model are based on the size of the smaller of the two sets. To test this, the data was grouped based on the size of the smaller set of the two used for the comparison.

The results shown below are for a small set of one ligand only.

The following figure shows the relation between the mean of to the size of the second set in the comparison (in the notation used before this plot shows , .



When fitting a linear model to mean the model has an R-squared value of 0.999.

The following figure shows the relation between the SD of to the size of the second set in the comparison (in the notation used before this plot shows , .

When fitting a linear model to SD the model has an R-squared value of 0.9868.

[Please explain why you are doing it. Do you suspect that their results are inaccurate? Can you correct them? The main theme so far was about the influence of the size of the smallest set. No you compare data with human only to all. How is it related? Please revise your summary accordingly.]

In Keiser et al. (2012) SEA predictions were tested across a number of ligands and proteins. In order to check if we are on the right track I calculated normalized z-scores (ZS) in the same manner they were calculated in Keiser et al. (2007). I then plotted a histogram of the values and marked in vertical lines the ZS for positive SEA predictions. [maybe you mean specific positive predictions that Keiser has tested?]

Each target’s ZS (such as Sertraline) is calculated against two target proteins groups, the first is human only, and the second is all organisms for which binding information is available.

The reason I’ve added this plot is I have a feeling we are dealing with two populations, the right most of ZS which belongs to ligands that will bind to tested targets and the left most with of ZS which belongs to ligands that will not bind to tested targets.

